Eden Andes, Jon Clancy, Teresa Phan
INLS 509 - Information Retrieval
Final Paper

Jon Clancy, Teresa Phan, and Eden Andes

INLS 509

Searching for Web Bias

Introduction

Today, when we have a query or a need, we ask Google, and accept the given results. Our intention is to examine whether we should accept that answer. We want to know if results are fair, and whether they can be manipulated. Are less privileged and underserved populations unfairly represented in search engine rankings? We will address issues of race, gender, and politics within the scope of social media recommender systems and search engine bias. Search engine results impact the way we see the world, we will attempt to bring awareness to the fundamental biases that impact the search engine.

From 1999-2004, Abbe Mowshowitz and Akira Kawaguchi of the City University of New York published a series of studies aimed at identifying and quantifying bias in search engines. Some of the articles were written for the scientific community, and others were for the public. Using different methods and testing a number of hypotheses, they consistently came to the same conclusion paraphrased nicely in their September 2002 ACM article: The only realistic way to counter the ill effects of search engine bias on the web is to make sure a number of alternative engines are available. According to the data gathered by NetMarketShare.com in November 2018, more than 74% of all search (and 84% of mobile search) was performed on Google (Mowshowitz & Kawaguchi, Bias on the Web , 2002). If we are to accept the repeated

conclusions of Mowshowitz and Kawaguchi, we exist in a world that may be strongly affected by biases reflected in and caused by search engines. It is our endeavor to identify some of these biases, their real world impact, and what is being done and can be done to combat them.

Cathy O'Neil's Weapons of Math Destructions gives the example of a case in 2007, Washington DC schools implemented an algorithm to identify the worst performing teachers in their district. They hired a consulting firm to map out quantification of successful teachers. In 2011, one (and presumably several) 2nd year teacher, generally regarded of exceptional, was laid off by the algorithm. She was flummoxed as were parents and colleagues. The most likely explanation seemed to be that her incoming 6th grade students had come in with surprisingly high standardized test scores in reading comprehension, that did not prove to reflect their skill level. Later, it was shown that many tests from the incoming schools included erasures and corrections that could suggest that teachers in that school may have been correcting the tests. While evidence was never conclusive, the significant decrease in test scores of students certainly suggested some amount of questionable action on the part of the 5th grade school. Ultimately, the exceptional student was given strongly positive references for a new job, and was told that the algorithm that laid her off was very complex and there was no way of knowing why; that without proof of malfeasance on the part of the 5th grade school. Weapons of Math Destruction demonstrates that the teacher was forced to prove beyond a shadow of doubt that the algorithm was using bad data, while the algorithm was not held to any scrutiny (O'Neil, 2016). In other words, the City of Washington DC manufactured a recommender system for good teachers and, much more specifically, the worst teachers in their system. It determined what information should go into the collection, and then let it run. Because a struggling school system could show

that they were directly addressing the problem, the algorithm was deemed a success. However, the year in question 206 teachers lost their jobs because this algorithm determined they were the worst teachers in one of the worst school districts in America, so it is likely that many of those teachers did not leave with glowing references, and that other competent teachers now needed to start a different career. Due to the administration's overall happiness with the system, and the fact that it was able to score some really easy political points, it remained in place and its function remained opaque.

The problem of bias in information retrieval is many-fold. We will attempt to examine the algorithm itself, which may reflect the worldview and beliefs of the programmers, will certainly know more personally about those who are either computer literate active contributors to the web and more demographically about those who do not have access to computers, are institutionalized, marginalized, or involved in some form of social services. We also want to look at the data that any given algorithm might mine, whether it is categorized fairly or not structured in any way, it may still reflect the worldviews of the authors. We will attempt look at the impact of advertising and financial motivation on behaviors. Finally, we will look at how user behavior can have an impact on any system that employs feedback or machine learning to improve its function. In this way, the decisions reflected by search engines and recommender systems can have a significant impact on our decision making process, choices made by individuals, by groups or organizations and even governments.

Among the many challenges of shining a light on bias as it relates to large scale, for-profit information retrieval systems is their proprietary nature. Companies like Google, Facebook, Microsoft, Amazon, and Apple do not want to let too many people have a look under

the hood, as it could lead to secrets being released. Even more closely guarded is the technology and research behind adversarial information retrieval, spam and malignant content detection, all of which works to avoid large scale content bias in favor of one particular politician, religion, or even hate group. We can infer some methods from research into controversial and malicious Wikipedia content, and sentiment analysis as it pertains to sarcasm in Twitter feeds and Amazon reviews, but this is peripheral at best. Finally, because the most widely used and influential information retrieval systems are publicly held companies that depend on consumer confidence and popularity to remain on top, an in-depth exploration of how bias affects their users has the potential to be very damaging, and in itself suggests that the researcher may be subject to their own confirmation biases, which no doubt fuels any reluctance to outside research. It is worth noting that these large companies may have performed their own internal systematic reviews, but the existence of such work or its findings remains unpublished. As a result all of the research we found looking at bias in information retrieval is either performed on systems that have been shelved, from the outside looking in, or on smaller, less impactful systems.

Explanation of Individual Problems & Context

Within the context of social media, bias arises as a result of the application's personalization algorithm. Algorithms act a the gatekeepers of information, and its purpose in social media is to curate a user's newsfeed or timeline. According to Shoemaker, et al. (as cited in Nechushtai & Lewis, 2018), algorithms have now taken over many of the tasks that humans used to control, such as selecting what stories or posts will be seen and how they are ranked. Thinking of a newspaper, articles on the front page are what the editor believes the newspaper's

readers should perceive as most important. Much like the layout of a newspaper, personalization algorithms are built to produce a timeline with posts and stories that are considered relevant to the user, and is ranked by importance according to the calculations of the algorithm and the user's previous interactions on the site. While a user interacts with their timeline, each click, like, and share not only influences what they'll see in their timeline in the future, but also the timeline of users that they interact with as well. Because information is constantly being created and shared, social media platforms use these algorithms to keep up with the mass amounts of content that humans alone could not manage.

In Engin Bozdag's article, "Bias in algorithmic filtering and personalization," he references Facebook's personalization algorithm, EdgeRank, which uses a technique that predicts what the user would be interested in based on other similar user's interests, also known as collaborative filtering. He explains that EdgeRank works by calculating the sum of edges, consisting of affinity scores between users, the weight of the edge type, for example a comment on a post "weighs" more than liking a post, and the measure of time when the content was posted (Bozdag, 2013). The EdgeRank algorithm uses all of this information to decide what users see on their timeline, what pages they are recommended to like, and what events they may be interested in. Although personalization algorithms like EdgeRank shows users what they are thought to want to see and interact with, we must remember that it is also excluding posts, comments, news, and events it thinks users may not like. What would be the effects if we saw some content that we may not necessarily like? Is it not better to have content and posts with values that challenge our own? Nechushtai, et al. makes note of this concern such that more and more timelines, whether it is in regards to news or for social media, are predominately being

curated by algorithms than editors, and that users prefer this method (Nechushtai & Lewis, 2018).

With the use of personalization algorithms, users are subject to confirmation bias, which the Encyclopedia Britannica defines it as the "tendency to process information by looking for, or interpreting, information that is consistent with one's existing beliefs" (Cassad, 2016). Because users are seeing posts, comments, and news articles that they are likely to agree with, they then interact with the content through likes, comments, and shares. This then feeds that interaction information back into the algorithm, which outputs more filtered content that reaffirms the user's likes and beliefs. An example of this would be the spread of fake news appearing on social media sites during the 2016 election. In the months leading up to the election, Buzzfeed reported that there were 8,711,000 interactions with the top 20 fake news stories on Facebook alone (Silverman, 2016). For users who already knew who they were voting for, they were likely interacting with these fake news articles, making their case for or against a candidate and using the articles to reaffirm their beliefs. Through the personalization algorithm, this interaction with fake news articles on Facebook would have made its way into their friends' timelines, who may have been on the fence about a candidate, until they experienced the influx of fake news that confirmed their initial thoughts of a candidate. Whether or not a user realizes, they give into confirmation bias by interacting with the content shown on their social media, and the algorithm keeps churning out content it knows you will interact with, much like a feedback loop.

The concept of the echo chamber or filter bubble is similar to the effect of a personalization algorithm. They can be described as the space where a person's thoughts and beliefs are amplified and reinforced. These concepts also subject users to confirmation bias in

the same way the personalization algorithm filters information that may be contrary to a user's belief, only to reinforce their values. Pizzagate, a coined hashtag from Twitter, was an outcome of an echo chamber, where people on social media sites, such as YouTube, 4Chan, and Reddit were contributing to rumors that Hillary Clinton was running a child sex ring and Comet Ping Pong was one of the venues that housed the children. After hearing and reading about Pizzagate from social media, a man from North Carolina went to Washington D.C. to investigate, and fired shots in the restaurant, threatening the staff to disclose the evidence of the child sex ring (Fisher, Cox, & Herman, 2016). While it is frightening to think that something like this could happen because of the spread of fake news, social media sites are setup to do just that - spread information, whether it is good or bad, real or fake.

Gottfried and Shearer (as cited in Sphor, 2017) referenced from the survey conducted by the Pew Research Center that 62 percent of adults in the US receive their news through social media platforms. With the effects of the echo chambers and filter bubbles within social media, people may be exposed to fake news such like Pizzagate, but with no countering information to tell users that this is not true. It also may be hard for users to discern what is true or not true if there are others, including their friends on social media, who may agree with these stories, and provide further corroborating "evidence" to the original post. Kate Turetsky and Travis Riddle found that environments that produce or contribute to the echo chamber effect can cause users to become more extreme in their beliefs (Turetsky, et al., 2017). As many users of any social media site has experienced, it is quite easy to go down a rabbit hole once something catches your eye, especially when there is plenty of information to supplement and reinforce those beliefs.

When social media sites like Facebook and Twitter were created, they were marketed as creating a platform brings users together from all over the world. Tom Wheeler writes that algorithms have actually siloed communities, preventing them from allowing the "democratic processes to succeed" (Wheeler, 2017). He goes on to discuss that what is good for a business isn't necessarily good for the people, such that because of the way social media structures its content, Americans were prone to exploitation, as indicated by the election meddling by people in Russia. Wheeler makes a convincing point that while we can point our fingers at the Russian meddlers, it is really the social media platforms that created, enabled, and exposed our weakness in giving into confirmation biases.

With all of these issues that social media algorithms have created, how should we go about resolving them? A few solutions that have been offered are to minimize the bad effects and improve upon the good parts of the personalization algorithm (Bozdag, 2013); use a combination of the personalization algorithm and a human factor to mitigate the personalized posts (Powers, 2017); and to create an algorithm that identifies the source of the information that is posted (Wheeler, 2017). Each of these authors recognized the significant power of algorithms that sift and filter through vast amounts of information, and rather than completely abandoning algorithms altogether, they see them as implementational tools to change the effects of social media for the better.

Bozdag's solution to minimize the harmful effects and expanding the beneficial effects goes into designing an algorithm with good morals (Bozdag, 2013). This can become an issue when trying to decide what is considered "good" or "bad". If algorithms are being coded to recognize "good" or "bad" content, is the algorithm still not susceptible to biases of humans and

what they consider good morals? It is a difficult case when deciding what the output of the algorithms should be, because social media sites still want their users to interact with their platform and keep them engaged, without seeming like the platform is curating a timeline with "wrong" or irrelevant content.

After the 2016 election, Facebook was accused of allowing the spread of misinformation on their platform, and decided to go back to a "hybrid approach" where editors also help curate the trending news section (Powers, 2017). Facebook was operating on their hybrid approach before, but was accused then of curating the news section with anti-conservative bias. Again there are difficulties, this time in deciding how far to go when discussing whether or not a news article is actually anti-conservative or if it is just liberal, and not "anti" anything. Even with the algorithm working alongside editors, misinformation or "bad" articles and posts may still slip through, or social media sites can still be accused of leaning one way or the other if someone from a political party is reported or banned for making a slighted comment. It is seemingly a better system that a purely algorithmic filter, but even then, the biases of the editors makes its way through someone's timeline.

Wheeler's solution to create a public interest algorithm that shows where the information, i.e. post or news article, comes from is interesting such that he is metaphorically pitting two algorithms against each other (Wheeler, 2017). Would people have been less likely to share fake news articles if they knew where they came from? It is difficult to say because even now, people pit CNN and Fox News against each other because they are stereotyped by political affiliation, where people of different parties may take one news channel as truth and the other as the spread of false information. Depending who creates the public interest algorithm, for example the

Federal Communication Commission (FCC) versus a social media platform creating their own public interest algorithm, people still may not trust the algorithm because of who it is "controlled by" or the opposite, blindly trusting the algorithm just because it is "controlled" by a specific source. It seems ingenious, but at the same time simple, to take an algorithm to expose another algorithm in a way that makes social media platforms more transparent.

Of the three solutions offered, Wheeler's solution is likely the most informative option, which could be a useful counter to vast amounts of useless and useful content being created every minute. Awareness of the source of a post or news article can be very helpful in cases of identifying fake news sites as well as giving the user a sense of autonomy when deciding whether or not to trust the information in front of them after seeing the identified source through a public interest algorithm. Also, with the public interest algorithm, it can be easier for social media sites to have transparency in understanding who, where, and when content is being created and distributed, as well as identifying sources of fake news and taking down groups of users sharing information contributing to harmful confirmation biases concerning politics, race and gender.

Matthew Kay defines stereotypes as "[beliefs] that individuals in a group...generally have one or more traits or behaviors"(Kay, 2015). Bias emerges as a reaction to exposure to various misrepresentations of stereotypes, often at the expense of marginalized groups. In the digital age, widespread exposure these misrepresentations can have dire consequences for those in the crossfire. Nearly three quarters of Americans used a search engine in 2012, with the majority choosing to use Google (Purcell, 2012). Despite these numbers, Google has denied responsibility for all instances of inappropriate search results in the past, though they validated the existence of

these issues by going back and correcting them. This coupled with a general belief and trust in the legitimacy and accuracy of search results perpetuates and amplifies negative stereotypes, especially toward marginalized groups. In this section, we argue that search results are biased against certain genders and races, and that developers behind search algorithms embed their bias into the system and create skewed results.

Issues pertaining to racial and gender bias in information retrieval are volatile. Cultivation theory, first applied to television commercials in the 1980s and 1990s, argues that portrayals in media can develop, reinforce, or challenge viewers' stereotypes (Potter, 1993). Studies done on query results from different search engines have shown that this theory is still relevant today when applied to 21st century media. In 2015, Matthew Kay and colleagues conducted a series of tests on search results depicting gender representations in different career fields. Their research addresses what Jahna Otterbacher refers to as "person perception": the task of making quick and accurate judgments about people as a result of constant exposure to many walks of life (Otterbacher, 2016). They evaluated their search results through four different lenses: stereotype exaggeration, systematic over/under representation, qualitative differential representation, and perceptions of occupations in search results. They found evidence of gender bias in each, with women being underrepresented, oversexualized, or pigeon-holed to traditionally female-dominated careers. Backlash to these results correlates to gender bias held by participants of the study. Searchers prefered results which matched their preconceived notion of which gender was traditionally associated with a specific career: female results in a traditionally male-dominated occupation were not considered accurate or desirable. This mirrors results from a study conducted by Otterbacher and colleagues in 2017. They evaluated gender

bias on a more general scale, looking broadly at "person perception" between the binary male/female spectrum. As a whole, they found that women are associated with "warm" qualities, such as "emotional", "expressive", or "open-minded", and men are associated with "agentic" qualities, such as "assertive", "determined", or "independent". In addition to these findings, a search on the word "person" retrieved "over twice as many photos of men/boys as compared to women/girls"(Otterbacher, 2017). When results came back that were atypical, reactions were negative: women with agentic traits were not considered as powerful as men; what's more, searchers were adverse to atypical women more than atypical men.

The implications of gender-biased search engines are twofold when race is also taken into account. In her seminal book *Algorithms of Oppression*, Safiya Umoja Noble delves deeply into how People of Color, Black women and girls in particular, are on the extreme end of victimization by search engines. She first encountered this in 2011, when a search query on "black girls" returned oversexualized and pornographic images and websites (Noble, 2018). After bringing her results into the limelight, there was an outpouring from people of marginalized society with claims of the same biased and offensive search results. The problem, she states, isn't that the algorithms are biased, but the developers behind them. She addresses the prevalence of cultivation theory in the 21st century, stating that "traditional misrepresentations in old media are made real once again online and situated in an authoritative mechanism that is trusted by the public: Google" (Noble, 2018). A search on "gorillas" in 2017 brought back the faces of People of Color, while a search on "unprofessional hairstyles for work" retrieved images of Women of Color and their natural hair. This constant exposure to digital microaggressions and "algorithmic oppression" toward black women can lead to long term health concerns.

In 2018, the New York Times released a report documenting the abnormally high infant mortality rate for Women of Color in the United States (Villarosa, 2018). Their conclusion was that the the “lived experience” of being a Black woman in America has lead to an infant mortality rate that is twice that of white infants. In addition to this, the phenomenon of the “self-fulfilling prophecy”, wherein an untrue preconception about someone manifests itself in that person and eventually becoming true, effects women's’ success rates in traditionally male-dominated occupations (Spencer, 1999). This can have a snowball effect, as web-content screeners can also be influenced by these biases in their choices in what gets censored and what doesn’t. Constant exposure to negativity regarding one’s gender identity, race, or lifestyle can lead to lower self-worth, diminished ambition, and can work to reinforce the the underlying ways that gender and racial bias are endemic in society. The source of these biases are written into the search algorithms by developers themselves. This was made apparent in 2017, when a Google employee, James Damore, released an “anti-diversity” statement, alleging that women were not genetically fit to be coders and developers (Matsakis, 2017).

When pinpointing the source of bias in search results, the responsibility falls to the developers again and again. Given the inanimate nature of computers and algorithms themselves, this makes sense. Computers cannot be racist or sexist; at their core, they are simple machines that execute what they are told to do. If Google or Bing retrieves something that the searcher considers offensive or inappropriate, it’s not the engines that are racist, but the the person writing the algorithms that serve as the engine’s creators and nurturers. Solutions to this bias have been presented and sought after by professionals in the field. Sheryl Sandberg’s Lean In Foundation, in collaboration with Getty Images, aims to correct bias against women in stock images available

on the internet. In 2016, Black Girls Code, an organization that promotes Women of Color to be active participants in the fields of computer science and engineering announced it was establishing residency in Google's New York City Offices. This was part of a larger, multi-million dollar initiative by Google to diversity its work environment and to "create a pipeline of talent into Silicon Valley and the tech industries" (Noble, 2018). Both Kay and Otterbacher stress the responsibility of developers to be aware of "the importance for designers, particularly those who build applications on top of search [application program interface]s, to consider the stereotypes conveyed through search results, and lay the groundwork for developing methods for the automatic detection of bias in [query results]"(Otterbacher, 2017).

The majority of solutions which address search engine bias hinge upon developer awareness. Diversity training, vetting, and periodic diversity check-ins have all been recommended (Kay 2016, Otterbacher 2017, Noble 2018). Kay also places emphasis on correcting gender proportions in search results to better mirror gender labor statistics reported by the Bureau of Labor and Statistics. Interestingly, Otterbacher places more responsibility on users, stating that users themselves should be cognizant of how search engines and algorithms work. They also argue that bias can counter bias: engineers should focus less on creating neutral algorithms and more on creating ones that actively work toward gender equality (Otterbacher, 2017). On another side of the spectrum, Noble addresses the lack of diversity employed by Google and Bing: "Black women are not employed in any significant numbers at Google… jobs at [major tech companies] that could employ the expertise of people who understand the ramifications of racist and sexist stereotyping and misrepresentation and that require undergraduate and advanced degrees in ethnic, Black/African, women and gender, American

Indian, or Asian American studies are nonexistent" as are any advanced and in depth ethics requirements for engineering and computer science programs (Noble, 2018). These solutions, however, all assume that unbiased search results will also generate profit for the company. Noble argues that as long as the pornification of Black women and girls generates hits and money for Google, they will continue to use biased algorithms (Noble, 2018). Developers and companies will only approach the issue of equality and restructuring their algorithms if the incentive for them to do so is strong enough.

Methods used to investigate bias in IR

Cathy O'Neil's Weapons of Math Destruction postulates a series of criteria to determine if an algorithm has the ability to create problems. Those criteria are Opacity, Damage, and Scale (O'Neil, 2016). Opacity: she asserts that the process of big data algorithms is proprietary, and therefore not transparent. It is driven by many market factors, some of which are obvious, some of which are themselves discovered by algorithms. But because the process is not explained, we have no way of knowing how our data is used and no way to protect ourselves. This can involve using neighborhood information, credit scores, rental histories, or anything else about us that can be quantified. Damage: she discusses how these algorithm driven decisions can have real-life impact and have the capacity to cause an individual damage. For example, an HR department might have a system that helps it determine whether to hire an applicant. She additionally points out that these types of hiring choices based on data usually only apply to lower level employees and that upper level management is usually put through a series of meetings and interviews to determine whether they get a position. In this way, the algorithms disproportionately affect less

educated, less qualified people, while college educated professionals are largely unaffected. Scale: finally, she looks at the scale of the algorithm in question. A damaging algorithm that is used on a small scale will typically reveal itself quickly after one or two unfair victims emerge, but something that affects the employment of thousands of people and can be responsible for mass layoffs may, as a result, improve efficiencies and be 'right' a certain percent of the time. Even if it is a high percentage, that could still mean hundreds or even thousands of people incorrectly categorized.

While much of the research we are examining has taken a quantitative approach to measuring bias, O'Neil and some others have taken a much more qualitative approach, offering a narrative 'post-mortem' of the impact that large scale algorithms (usually recommender systems) on human behavior. For example, she cites the invention of the US News & World Report ranking of top universities. This was done in an effort to help guide high school students to the most enriching future they could have based on a series of metrics. The resulting popularity meant that low ranking schools changed their behavior to better fit the criteria. Low selectivity schools, historically known as safety schools to top high school students, scored poorly in the rankings in part because of their high acceptance rate. The high acceptance rate had been because they knew that a number of their top applicants would also get in somewhere else and only a small percentage of them would attend. Therefore, they developed an algorithm to identify students that would likely not attend the school and began rejecting those top applicants, thereby eliminating the reality of the safety school and possibly missing out on some excellent students. The result, however, was that they could show that they were more selective and thus rise in the

rankings. In this way, data that favors one seemingly preferable attribute is not comprehensive and has real-life consequences that in turn reinforce the data.

Similarly, Ricardo Baeze-Yates penned an article entitled 'bias on the web' which discusses broader concepts surrounding bias. Although there is a section of the paper called measuring bias, it is there primarily to describe what bias is and how challenging it can be to actually measure it. He outlines the various sources of bias that can impact the way any given algorithm can understand data on the web. Activity bias exists because, although anyone with access can freely add content to the internet, only a very small percentage (.04-4%) of internet users are actually contributors. This number drops dramatically when you measure content that is deemed in some way useful. Access to the internet creates its own biases, as only about half of the world's population has access to the internet, only a portion of those with access are competent in digital media, and more than half of the internet is written in english. Additionally, domain trust decreases when the domain in question is not .com, .edu, .gov or .org, domains typically associated with the US, so it is likely to have a lower ranking on Google. Web spam increasingly contributes to the overall measure of biased information on the internet. Although companies continue to work to detect, filter, and combat web spam, an arms race exists and the spammers keep getting sneakier. Algorithms that deal with biased content, in and of themselves, do so based on the instructions (and therefore the biases) of their developers. The example given is that a user might want to read the news. Most of the news generated on any given day comes from governments, large companies and population centers. If the user lives in a small town, there probably isn't a great deal of news or popular news being generated. Therefore the algorithm either has to prefer local over popular news, or popular news over local news. Part of

the solution, as with many recommender systems, is to simply ask the user, but even when this is done, we run into issues of presentation and selection bias. Presentation bias is the that, from your perspective, the only news that you know exists is the news that you have been exposed to, and the choice of how to expose what information is deeply challenging. Newspapers have known for over a century that putting an article below the fold would mean some people wouldn't see it. Selection bias occurs on many levels, but includes the fact that users disproportionally click on the first result in Google, and are also more likely to click on information they agree with even when an overwhelming number of results might suggest that a competing viewpoint is more popular, relevant or likely to be true, confirmation bias. He also outlines the ways in which the desire to monetize a system can bias the behavior of a developer. In other words, ads diminish user experience, Therefore, improving user experience can be seen as a way to allow you to show more ads without ultimately decreasing user experience. In this same way, the varying biases of the data, the developers, the users and the interface all reinforce themselves creating vicious cycles further served by things like filter bubbles that select the information users have declared they want to hear, which have a tendency to undermine diversity, serendipity, novelty and opposing viewpoints. Baeze-Yates argues that the complex interworking of all the various biased behaviors of users and algorithms creates a deeply complex problem, and one that is very hard to account for (Baeze-Yates, 2018).

Mowshowitz and Kawaguchi took a far more quantitative approach, conducting a series of experiments in which they performed searches on 16 popular web search engines (most of which are marginalized or no longer in existence) to identify variety in results and rankings and to measure when the variety seemed to prefer one viewpoint over another. They were able to

demonstrate the existence of bias across several platforms and no 'completely fair' resource emerged. They searched for terms that, by their nature, demand 'for' or 'against' positions, and where the merits for and against could be argued to be comparable. They then measured the number of 'for' and 'against' positions represented by the top 16 search engines. Using the most commonly generated results, they calculated a response vector for all search engines. In this way, they could use Vector Space Models and cosine similarities to calculate deviation from the norm. Challenges include determining what search engines should be included to establish a reasonable baseline vector for a given query as this could affect the results given that some of the search engines and websites returned by search engines were no longer in operation at the time of publication, and they surmised that new search engines could emerge and gain popularity. The problem they identified is to determine whether the norm or the outlier is less biased, 'the results may mean that the engine picks up interesting items not found by the others (Mowshowitz & Kawaguchi, Measuring search engine bias, 2005).' In other words, this deviation from the norm could indicate a bias toward antiquated or malicious content, but just as easily toward serendipity or novelty. It could also simply indicate that the engine was not picking up relevant documents. The two most significant findings were that results vectors give us another way to assess search engine results in addition to precision and recall, and that the most important way to mitigate bias on the web is to have variety in search tools (Mowshowitz & Kawaguchi, Measuring search engine bias, 2005).

How are results being evaluated?

There is no question that human behavior and decision making is deeply affected by our inherent and learned biases. The idea that a person can make an important decision about another person that is completely objective is unrealistic, and in many cases we actually want those biases to be part of the decision (eg. I want to hire people I think I will enjoy working with, even if the candidate I don't enjoy is more qualified). In this way, it is important to note that bias is not inherently good or bad. We are preprogrammed to prefer 'us over 'them', where we draw that line may determine how much our biases are problematic to society (eg. the line could mean that a person is sexist, racist, xenophobic, or just afraid of bears). There is a long list of identified cognitive biases, many of which affect the way we interact on the web and respond to the results of a query. An example is the anchoring bias which leads us to prefer the first piece of information we are given over subsequent information, even when that information is called into question, this can be directly linked to users overwhelming preference of the first 1-2 results on a Google search and their influence on a person's beliefs. There is also the confirmation bias which would lead a person to prefer the article that conforms to their currently held over any other results. Others biases might affect the way we engage or contribute to the web, preferring to sharing posts (memes are particularly good at this) that make your own position sound smart and opposing positions sound simple or heartless. So there is a major portion of the bias existing on the web that can be attributed to the individual user.

There is also the question of access to the web. About half of the world has internet access. Among them, a significant portion is not tech savvy, this group disproportionally includes older and poorer populations. Much of the internet is english, while only about 20% of the world's population speaks english (Lyons, 2017). Much of what is known about internet use

along with behavior and beliefs of users is drawn from a pool of users that do not reflect the human population. So the data about users that drives the behavior of any search engine is based on an incomplete sample.

Finally, the algorithms themselves are biased. Whether this is because they reflect the biases of the developers, the biases of the users, the motivations of the parent company, or some component of the machine learning they use to refine themselves is incredibly hard to calculate. We do know that when a search engine that depends on ad revenue (like Google) receives a query, it runs two simultaneous searches, one through the web as it has indexed it, and the other through its collection of advertisements to determine whether there is a relevant ad to provide to users. While this is the understood status quo, it certainly constitutes some level of conflict of interest and, although they have clearly labeled paid advertisements as such, they are still subject to the preferential biases that users show for the first result on a list. Additionally when the second result beneath the ad is the same website, it still emphasizes cognitively reinforces the relevance of that result to a user.

The question remains, however, what do we do? First, there exists a comprehensive list of fairness measures (fairnessmeasures.org) that can be applied to computing tools, attempting to offer a set of standards that can help to combat bias. It will become increasingly critical to establish checks and balances, whether they are these standards or ones with similar intent as a part of the future of information retrieval and of all data science. Second, there is a need for awareness, both on the part of developers, but also on the part of the public, that the internet is not a repository of knowledge; it is a living breathing collection of thoughts, feelings, ideas, and resources, some of them well intended, some less altruistic. Third, some measure of transparency

for the companies like Facebook and Google (among others) that serve as information utilities, bringing the content of the web to millions of users per minute and having measurable impact on widespread beliefs and human behavior (eg. Facebook's impact on US elections has been widely documented).

The opportunity of the internet is to bring people together in ways not possible 30 years ago, to bring knowledge around the world and to build connections of knowledge and interests. We can generate massive amounts of data that reflect transactions, medical treatments, connections, and measurements on a scale that is almost impossible to imagine. The power of this data is largely untapped, but we will be retrieving information from it using algorithms. These algorithms will drive the way financial, political, environmental and ethical decisions are made around the world, and they will affect people who are not even contributing to this boon of data. It will be the responsibility of those who venture into the various fields of data science to consider ethics, fairness, and bias as foundational components of the powerful tools we are building, and to consider and mitigate the harm that a faulty system could cause. It will further be our responsibility to examine the situations in which these systems do cause harm, to learn from them, and build on those lessons.

Bibliography

Baeza-Yates, R. (2018). Bias on the Web. *Communications of the ACM*, 54-61.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethic and Information Technology, 209-277. Retrieved from UNC Libraries: https://doi-org.libproxy.lib.unc.edu/10.1007/s10676-013-9321-6

Casad, B. J. (2016, August 1). Confirmation Bias. Retrieved from Encyclopedia Britannica: https://www.britannica.com/science/confirmation-bias

Fisher, M., Cox, J. W., & Hermann, P. (2016, December 6). *Pizzagate: From rumor, to hashtag, to gunfire in D.C.* Retrieved from The Washington Post: https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html?utm_term=.d458de55fc40

Kay, M., Matuszek, C., & Munson, S. (2015). Unequal representation and gender stereotypes in image search results for occupations. Paper presented at the *, 2015*- 3819-3828. doi:10.1145/2702123.2702520

Kurtzleben, D. (2018, April 11). Did Fake News On Facebook Help Elect Trump? Here's What We Know. Retrieved from National Public Radio, Inc: https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election

Matsakis, L. (2017). Google Employee's Anti-Diversity Manifesto Goes "Internally Viral." Motherboard. Retrieved from www.motherboard.vice.com.

Lyons, D. (2017). *How Many People Speak English, And Where Is It Spoken?* Retrieved from Babbel Magazine: https://www.babbel.com/en/magazine/how-many-people-speak-english-and-where-is-it-spoken/

Mowshowitz, A., & Kawaguchi, A. (2002). Bias on the Web . *Communications of the ACM*, 56-60.

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing and Management*, 1193-1205.

Nechushtai, E., & Lewis, S. C. (2018). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. Computers in Human Behavior, 298-307.

Noble, S. U. (2018). Algorithms of oppression : how search engines reinforce racism. New York: New York University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown. Audio Book

Otterbacher, J. (2016). New Evidence Shows Search Engines Reinforce Social Stereotypes. Harvard Business Review. Retrieved from www.hbr.org.

Otterbacher, J., Bates, J., & Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. Paper presented at the *, 2017*-6620-6631. doi:10.1145/3025453.3025727

Potter, W. (1993). cultivation theory and research - a conceptual critique. Human Communication Research*, 19*(4), 564-601.

Purcell, K., Brenner, J., and Rainie, L., (2012). Search Engine Use 2012. Pew Research Center. Retrieved from www.pewinternet.org.

Silverman, C. (2016, November 16). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. Retrieved from Buzzfeed News: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.kqE3d8d5W

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4-28. doi:10.1006/jesp.1998.1373

Spohr, D. (2017). Fake news and ideological polarization Filter bubbles and selective exposure on social media. Business Information Review, 150-160. Retrieved from UNC Libraries: https://journals-sagepub-com.libproxy.lib.unc.edu/doi/full/10.1177/0266382117722446?utm_source=summon&utm_medium=discovery-provider

Villarosa, L. (2018). Why America's Black Mothers and Babies Are in a Life-or-Death Crisis. The New York Times. Retrieved from www.nytimes.com.